Phone Duration Modeling Using Clustering of Rich Contexts

Tanel Alumäe, Rena Nemoto

Institute of Cybernetics at Tallinn University of Technology, Tallinn, Estonia

{tanel.alumae, rena.nemoto}@phon.ioc.ee

Abstract

This paper describes a phone duration model applied to speech recognition. The model is based on a decision tree that finds clusters of phones in various contexts that tend to have similar durations. Wide contexts with rich linguistic and phonetic features are used. To better model varying and non-stationary speaking rates, the contextual features also include the observed duration values of previous phones. For each resulting phone cluster, a log-normal distribution of duration is estimated. The resulting decision tree and the log-normal distributions are used to calculate likelihoods of phone durations in N-best lists. Experiments on two Estonian recognition tasks show a small but significant improvement in speech recognition accuracy.

Index Terms: duration modeling, speech recognition, decision trees

1. Introduction

The weaknesses and limitations of hidden Markov models (HMMs) in modeling the variability of phone durations in speech recognition are well known. HMMs model phone durations using state transition probabilities which result in geometric probability density functions (PDFs) for phone durations, which inadequately represent the temporal behaviour of speech. Several methods have been proposed to resolve this limitation. Improved duration modeling can be integrated directly into the HMM framework, by replacing the HMM state transition probabilities with explicit duration PDFs [1] or by modifying the HMM topology [2]. However, these approaches significantly increase the computational complexity of decoding and offer no or only limited decrease in word error rates. Therefore, an alternative approach is to model word or phone durations using an independent model and use it as a separate knowledge source during N-best rescoring [3, 4] or, more recently, lattice rescoring [5, 6, 7]. Such approaches usually result in a small (usually not more than 5% relative) word error rate reduction in large vocabulary continuous speech recognition (LVCSR) tasks.

Our interest in duration modeling is partly inspired by the application of LVCSR for Estonian. Estonian is a quantity language with three degrees of length: short (Q1), long (Q2), overlong (Q3). Different quantity degree manifests phonological distinction. For example, the phoneme /a/ in the word *kalu/kaalu* can be realized as Q1 (/kalu/, 'fish', partitive plural), Q2 (/kaalu/, 'scale', genitive singular) or Q3 (/kaa:lu/, 'scale', partitive singular). The distinction between quantity degrees is not realized by the absolute length of the corresponding phone: the key to differentiate between Q1, Q2 and Q3 words is rather the ratio between the durations of phones in the first stressed syllable and the duration of the vowel of the following syllable [8]. Most of studies about the Estonian quantity system have investigated perceptual experiences and/or speech analysis [9, 10, 11].

In this paper, we apply a phone duration model for N-best rescoring in two Estonian LVCSR tasks. The model has some key differences from previously proposed approaches to phone duration modeling. First, we do not estimate explicit duration models for individual phonemes or triphones: rather we use phonetic and linguistic features of phones and their neighbouring phones to create phone clusters that have similar durations in the training corpus. This allows us to implicitly model effects such as pre-pausal lengthening which must otherwise be modeled using additional effort. Secondly, we use the duration values of previous phones as contextual elements when calculating phone likelihoods. This allows the model to use phone duration ratios for distinguishing Estonian quantity degrees, but also to take into account local speaking rate. Apart from the implementation of extracting linguistic and phonological features, the model is not specific to Estonian.

The following section describes the duration modeling approach used in this paper. Section 3 presents evaluation setup and experimental results. Section 4 concludes and discusses future work.

2. Modeling approach

2.1. Model design

With explicit duration models, the speech recognition problem is usually reformulated as a task of finding the best word sequence $W^\star$ and the corresponding word durations $D^\star$, given the acoustic signal. Assuming that, given the word sequence W, acoustic features A can be viewed as conditionally independent from word durations D, we can write:

$$\begin{aligned} W^\star, D^\star &= \underset{W,D}{\operatorname{argmax}} P(W, D|A) \\ &= \underset{W,D}{\operatorname{argmax}} P(A, D|W)P(W) \\ &= \underset{W,D}{\operatorname{argmax}} P(A|W)P(D|W)P(W) \end{aligned}$$

The third line in the last equation relies on the (invalid) assumption that A is independent from D, once conditioned on W. Therefore, the task of the explicit duration model is to estimate the likelihood $P(D|W)$. Note that this assumption fails to take into account the influence of the intrinsic duration probability density functions associated to the state-to-state transition probabilities of the HMMs of the acoustic model that are already factored in $P(A|W)$ [6].

Most previously proposed explicit duration models (e.g., [12, 4, 6, 7]) decompose the calculation of $P(D|W)$ using the likelihoods of word durations $P(d_i|w_i)$:

$$P(D|W) = P(d_i, ..., d_k|w_i, ...w_k)$$

In order to make the computation of this probability feasible, the duration models are approximated by considering each word

duration independent of the other durations in the utterance. However, to account for the natural variances among speaking styles and speakers, overall average phone duration is almost always normalized by the speaking rate of the utterance or speaker. Then, the probability durations of normalized word durations are approximated as follows:

$$P(d_i, ..., d_k|w_i, ...w_k) \approx \prod_i P(d_i^{(norm)}|w_i)$$

Our approach is to decompose the duration structure D into m phone durations $d_i^{(p)}$:

$$P(D|W) = P(d_1^{(p)}, ...d_m^{(p)}|W)$$

The likelihoods of phone durations can be decomposed using the chain rule and approximated using the Markov assumption:

$$\begin{aligned} P(d_1^{(p)}, ...d_m^{(p)}|W) &= \prod_{i=1..m} P(d_i^{(p)}|d_1^{(p)}, .., d_{i-1}^{(p)}, W) \\ &\approx \prod_{i=1..m} P(d_i^{(p)}|d_{i-n-1}^{(p)}, .., d_{i-1}^{(p)}, W) \end{aligned}$$

That is, we approximate the likelihoods of phone durations by conditioning on the durations of the previous n phones and the hypothesized words. By conditioning on the durations of the previous n phones we want our model to learn the effect of the (local) speaking rate on the expected phone duration. Furthermore, this approach allows us to capture speaking rate variations within an utterance, as opposed to normalizing over the average speaking rate of the utterance.

We use binary features to encode the dependence of durations $d_i^{(p)}$ on the word sequence W. The features look at the type and properties of the phone that the duration corresponds to, the position of the phone in the word and utterance, and features of the neighbouring phones. To condition the duration probability on the previous durations, we use the duration values of previous phones simply as discrete-valued features (representing the number of feature frames or milliseconds). The full feature set used in our experiment on Estonian speech is summarized in Table 1. Note that we do not try to model the duration of silence and filler units, but we do use their features (including their duration) as context elements.

We are thus left with a task of building a model for estimating likelihoods of phone duration, based on a set of binary and discrete features $x_i^{(p)}$ for that phone. In other words, our training data consists of pairs $(x_i^{(p)}, d_i^{(p)})$, and our goal is to build a model for estimating $P(d_i^{(p)}|x_i^{(p)})$ for unseen $(x_i^{(p)}, d_i^{(p)})$.

Our proposed duration model builds a binary decision tree regression model on the training data, using the mean squared error as the objective function. The tree is built using a fairly large minimum number of samples required for each tree node (we found the value of 100 to be optimal). The resulting tree leaves represent such feature subsets that tend to have similar durations. From the training samples corresponding to each leaf, we estimate a log-normal distribution of phone duration. At test time, we look up the cluster corresponding to the phone occurring in a recognition hypothesis using the decision tree, and calculate the likelihood of the observed phone duration using the log-normal distribution corresponding to this cluster.

2.2. Model introspection

We used the described approach to build a decision tree on Estonian phone alignments from approximately 140 hours of speech data from various sources (see next section for details). One way to understand and interpret the discriminatory power of different features is to calculate feature importance. In regression trees, the importance of a feature is computed as the (normalized) total reduction of the mean squared error brought by that feature (also know as Gini importance). Table 2 shows 10 features with the highest Gini importance. We see that the highest-scoring feature is a binary indicator of the next phone being a silence, i.e., it asks whether the current phone is pre-pausal. This is of no surprise, since the effect of pre-pausal lengthening is well-known and in previously proposed duration models, the effect of pause context is often explicitly modeled (e.g., [12, 4]). Besides other expected features, such as the phonetic length, existence of pause or utterance boundary, the highest scoring features also include the observed durations of the previous phones. This suggests that the decision tree is able to take advantage of the durations of previous phones when creating phone context clusters.

The decision tree is common for all different phonemes. The exact identity of the phone is simply modeled using binary features. It turns out that such features do not have particularly high importance: the highest-ranking phoneme identity feature belongs to the phoneme /j/ at the 27th position. This can be explained by the fact that the palatal approximant /j/ has almost always a very short duration in Estonian.

Figure 1 shows the top nodes of a trained decision tree. The text given with the non-terminal nodes corresponds to the binary question of that decision tree node. For each node, the graph also shows the mean duration value of all the training data pass-

Table 1: Features used by our duration model. All features except the last ones are binary.

Feature type	*#Features*
Phoneme	27
Phonemic type (vowel, consonant, nasal, etc)	9
Phonetic length (extra short, long, overlong)	3
Word boundary (before or after)	2
Current syllable position in word	10
Subtotal	51
Features of 2 previous phones	102
Features of 2 next phones	102
Utterance boundary	4
Duration of 2 previous phones (discrete)	2
Total	261

Table 2: Features with the highest Gini importance.

Feature	*Importance*
$phone_{i+1}$ = silence	0.4900
$phone_i$ is phonetically long	0.3300
$duration_{i-1}$	0.1200
$phone_{i-1}$ = silence	0.0270
$phone_i$ is vowel	0.0082
$phone_i$ is utterance-ending	0.0035
$duration_{i-2}$	0.0027
$phone_{i+2}$ = silence	0.0020
$phone_i$ is utterance-starting	0.0014
$phone_i$ is stop consonant	0.0011

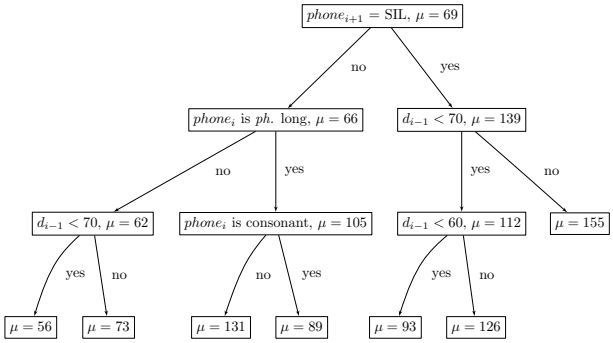


Figure 1: Top nodes of a trained decision tree, with questions and mean duration values corresponding to the nodes.

ing through that node, given in milliseconds. The mean duration over all phones is 69 ms (top node). If the phone is pre-pausal, the mean duration doubles, and increases even further to 155 ms if the duration of the previous phone is at least 70 ms. Another interesting observation from this tree fragment is that the mean duration of phonetically long (and not pre-pausal) consonants is noticeably shorter (89 ms) than for long vowels (131 ms).

Another way to visualize the duration model is to look at the terminal nodes (i.e., certain phone clusters) of the decision tree. Figure 2 shows the differences of randomly chosen word-internal and sentence-final phone clusters. The figures show an empirical duration distribution for a cluster (unfilled histogram, in hundredths of seconds), the corresponding log-normal distribution probability density function (PDF, solid line), the log-normal duration PDF over all training data (dotted line), and the test data duration histogram for that cluster (filled histogram). As expected, the PDFs are quite different: the PDF of the word-internal phone puts much more probability mass on shorter durations, while the PDF of the sentence-final cluster is much flatter, with more probability mass on longer durations.

We evaluated, how much each feature set contributes to performance of the model. To do this, we compared the duration models with different feature sets, using development set perplexity as the performance measure. Perplexity of the model on N observed durations of the development data is defined as:

$$\mathrm{PPL} = e^{-\sum_{i=1}^{N} \frac{1}{N} \ln \hat{p}(d_i|x_i)}$$

Table 3 shows how the perplexity of the model decreases when richer features are added to the model. Perplexity consistently drops with more features added until the context of four previous phones instead of two is used, in which case model performance slightly degrades. Using the durations of only two previous phones also resulted in best results in speech recognition experiments, as increasing the context to four preceding phones didn't improve the rescoring results.

Table 3: Perplexity (PPL) of the duration model with various feature sets (lower scores are better).

Feature set	*#Features*	*PPL*
Phone identity, phone type	36	10.6
+ Phone phon. length, word and utterance boundary, syllable number	55	9.2
+ Features of two previous phones	157	8.5
+ Features of two next phones	259	7.9
+ Durations of two previous phones	261	7.1
+ Features and durations of four previous phones instead of two	365	7.2

3. Speech recognition experiments

3.1. Speech data

For training the acoustic models (AMs), we used various wideband Estonian speech corpora, totalling in about 140 hours:

- the BABEL speech database which contains about 9 h of dictated speech;
- a corpus of Estonian broadcast news which contains mostly dictated speech, with some semi-spontaneous studio and telephone interviews (36 h);
- a corpus of broadcast conversations, consisting of various semi-spontaneous talk shows from three radio stations (20 h);
- a corpus of semi-spontaneous (mainly telephone) interviews from TV and radio programs, discussing mainly daily news and current events (18 h);
- a corpus of local conference speeches and university lectures, recorded with a close-talking microphone (40 h);
- a corpus of studio-recorded spontaneous monologues and dialogues (16 h)

For tuning and measuring system performance, two different domains were used: broadcast news (BN) and broadcast conversations (BC), with separate development and evaluation sets for both of the domains. The development and evaluation sets for BN consist of 30 short radio news programs with mainly dictated speech, both around two hours in total. The development and evaluation set for the BC domain contain spontaneous TV and radio interviews, both around one hour in total.

3.2. Speech recognition system

Details of the Estonian transcription system are described in [13]. The acoustic models are continuous triphone HMMs with 2000 Gaussian mixtures that use 385 150 Gaussian distributions. Tied-state cross-word triphones estimated using maximum likelihood training are used to model 25 phonemes and silence/noise and garbage. LDA-transformed MFCC features are used.

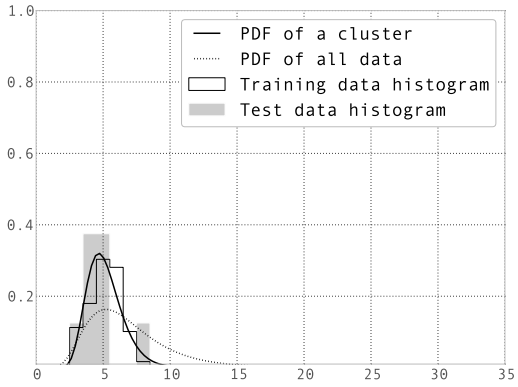
Language model training data contains around 300M words from various sources (newspapers, popular magazines, web, parliament transcripts, fiction, manual transcriptions of broadcast news, broadcast conversations and lectures).

As Estonian is a heavily compounding and inflective language, the lexical variety of the language is very high. To reduce the out-of-vocabulary (OOV) rate of the LM, compound words are decomposed into compound particles, using the word structure information assigned by a morphological analyzer [14]. After decoding, compound words are reconstructed from the recognized particles using a hidden event LM.

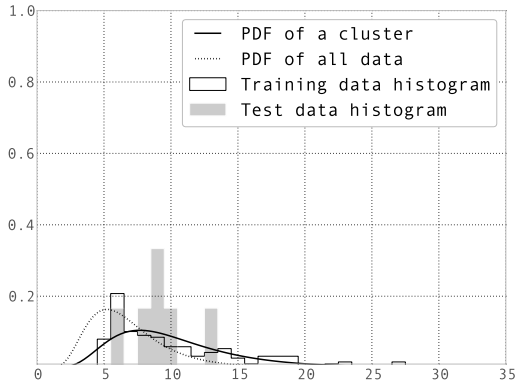
We use a vocabulary size of 200K units. For both of the domains used in the experiments, we created an optimized 4-gram LM by interpolating source-specific LMs estimated using interpolated modified Kneser-Ney discounting. A rule-based system was used for deriving the pronunciations for words in the LM lexicon.

3.3. Duration models

We used the 140 hour speech corpus (the same as for estimating acoustic models) for training the decision tree based dura-



(a) Word-internal phone cluster.



(b) Sentence-ending phone cluster.

Figure 2: Training data histogram, the resulting PDF, test data histogram and the PDF over all training data for a randomly chosen word internal phone cluster (a) and sentence-ending phone cluster (b).

tion model. The speech corpus was aligned with the reference transcripts using the AM trained from the same corpus. This resulted in training data of 4 902 618 phones. Using the features of the phones and their corresponding reference durations, we built a decision tree using the Scikit-learn toolkit [15]. We tried different values for the minimum number of samples required for each tree leaf, and chose the value of 100 after optimizing the likelihood of development data. The resulting tree had 35 581 leaves. The same model was used for both domains in the recognition experiments.

We also experimented with some more advanced model configurations. We tried estimating a log-normal *mixture* model for each phone cluster, and building a random forest instead of a single decision tree. Both of those approaches failed to decrease the model perplexity on development data.

3.4. Decoding and rescoring

Speech in the development and evaluation sets was decoded using a carefully optimized baseline system [13]. For each segment, we generated a 100-best list of recognition hypotheses. All hypotheses were aligned with the corresponding speech to obtain phone duration timings. The trained duration model was then used to produce likelihoods of non-filler phone duration for all hypotheses. Likelihoods for individual phones were multiplied to obtain likelihoods for the whole hypothesis. An extra constant proportional to the number of phones in the given hypothesis was produced for N-best lists, to serve as a phone insertion penalty. Combination weights for all scores (AM, LM and DM likelihoods, number of words in the hypothesis, number of phones is the hypothesis) were then independently optimized on both of the development sets for word error minimization. Simplex-based "Amoeba" search implemented in SRILM [16] was used for optimization. To have a fair comparison, we also optimized the score combination weights for the baseline system. Finally, the optimized weights were used to obtain new baseline and duration score augmented recognition hypotheses.

The resulting word error rates are reported in Table 4. The use of duration models reduces the word error rate by 1.4% absolute (4.3% relative) for the BC evaluation data and by 0.3% absolute (2.2% relative) for the BN evaluation data. The gains are relatively small, but for both sets, they turned out to be statistically significant (according to the Wilcoxon test).

Table 4: Word error rates for two Estonian transcription tasks before and after rescoring with the duration model.

	BC		BN	
	Dev	Eval	Dev	Eval
Baseline	37.2	32.4	13.5	13.8
+ Duration model	36.5	31.0	13.2	13.5

4. Conclusion

This paper described a method for estimating and applying phone duration probabilities in speech recognition. We use a decision tree based regression model to group phones into clusters that have little deviation in duration, based on phonological and linguistic properties of phones and several neighbouring phones. To accommodate the effect of varying and non-stationary speaking rate, we use the duration values of previous phones as additional features during clustering. The use of such rich contexts eliminates the need for building explicit models for phones in pre-pausal context and helps us to avoid normalizing the duration values based on explicitly calculated speech rate.

We tested the proposed model on two Estonian speech recognition tasks. The use of duration models in N-best rescoring resulted in small but statistically significant reductions in word error rate. The improvements are similar to those that have been obtained with more complex duration models.

Future work includes applying the model to speech recognition lattices, experimenting with other languages and adapting the model to individual speakers.

5. Acknowledgements

This research was supported by the European Regional Development Fund (ERDF) through the Estonian Center of Excellence in Computer Science (EXCS) and the Estonian Ministry of Education and Research target-financed research theme No. 0140007s12.

6. References

[1] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Compututer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[2] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *ICASSP 1985*, 1985.

[3] A. Anastasakos, R. Schwartz, and H. Shu, "Duration modeling in large vocabulary speech recognition," in *ICASSP 1995*, vol. 1, 1995, pp. 628–631.

[4] V. R. R. Gadde, "Modeling word duration for better speech recognition," in *Speech Transcription Workshop*, Maryland, 2000.

[5] D. Povey, "Phone duration modeling for LVCSR," in *ICASSP 2004*, vol. 1, 2004, pp. 829–832.

[6] D. Seppi, D. Falavigna, G. Stemmer, and R. Gretter, "Word duration modeling for word graph rescoring in LVCSR," in *Interspeech 2007*, 2007, pp. 1805–1808.

[7] N. Jennequin and J.-L. Gauvain, "Modeling duration via lattice rescoring," in *ICASSP 2007*, vol. 4, Honolulu, HI, USA, 2007, pp. 641–644.

[8] A. Eek and E. Meister, "Simple perception experiments in Estonian word prosody: foot structure vs. segmental quantity," in *Estonian Prosody*, Tallinn, 1997, pp. 71–99.

[9] P. Lippus, "The acoustic features and perception of the Estonian quantity system," Ph.D. dissertation, Tartu University, 2011.

[10] P. Lippus, E. L. Asu, P. Teras, and T. Tuisk, "Quantity-related variation of duration, pitch and vowel quality in spontaneous Estonian," *Journal of Phonetics*, vol. 41, no. 1, pp. 17–28, 2013.

[11] E. Meister and L. Meister, "Native and non-native production of Estonian quantity degrees: comparison of Estonian, Finnish and Russian subjects," in *Nordic Prosody*, Tartu, 2012, pp. 1–9.

[12] G. Chung and S. Seneff, "Hierarchical duration modelling for speech recognition using the ANGIE framework," in *Eurospeech 1997*, 1997, pp. 1475–1478.

[13] T. Alumäe, "Transcription system for semi-spontaneous Estonian speech," in *Baltic HLT 2012*, Tartu, Estonia, 2012.

[14] H.-J. Kaalep and T. Vaino, "Complete morphological analysis in the linguist's toolbox," in *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, Tartu, Estonia, 2001, pp. 9–16.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Interspeech 2002*, Denver, Colorado, USA, 2002.